RESEARCH ARTICLE OPEN ACCESS

Estimation of Rainfall Missing Data in an Arid Area using Spatial and EM Methods

Zeinab Abu Romman*, Jawad Al-Bakri**, Mustafa Al Kuisi***
**(Department of Land, Water and Environment, The University of Jordan, Jordan.*
*** (Department of Land, Water and Environment, The University of Jordan, Jordan.*
****(Department of Geology, The University of Jordan, Jordan.*
Corresponding Author: Zeinab Abu Romman

ABSTRACT

Outputs from hydrological studies are dependent on rainfall data that is controlled by the limited spatial distribution of stations, interrupted time series of data, variable rainfall among locations in the area under consideration. Therefore, it is essential to fill the missing records using different methods. This study aims to infill the missing rainfall data for an arid area located in north Jordan using the inverse distance weighted (IDW) and the expectation maximization (EM). Different statistical tests were used to assess the accuracy of both methods. Results showed that EM method results were highly correlated with the raw data, and show a perfect similarity result with a root mean square error (RSME) of less than 11 mm. Estimates were better for the station of Mafraq than for the two other stations used in the study. This was attributed to the low percent (2.8%) of missing data for this station. The reason behind the differences in accuracy of estimates were attributed to the nature of the method and the percent of missing data for each station.

Date of Submission: 21-02-2019 Date of acceptance: 10-03-2019

I. INTRODUCTION

Rainfall data is important for hydrological modeling, agricultural and water budget estimation. Although the causes of rainfall variability may come from more global phenomena [1], it is important to know and understand their historical patterns to enhance decision making in arid and semiarid watersheds [2,3,4]. Therefore, for performing the effective rainfall analysis, it is essential to estimate the missing value in rainfall series. For that purpose, different methods are used for estimating the missing rainfall data for specific regions.

Methods for filling missing rainfall data include statistical and spatial methods. Statistical methods substitute missing values by the series mean method [5]. Thus, the average value of the series is not altered, and its variance is reduced [4]. Other statistical methods include the linear interpolation, which is appropriate for short temporal scales and variables with high autocorrelation, and the linear trend which adapts the variation of simple linear regression equations to ensure preservation of characteristics of the statistical parameter of the infilled data series [4,5]. More advanced methods are also proposed for the same purpose. An example on these methods is the expectation maximization (EM) algorithm, which is based on estimated regression models between missing and available data to produce the lowest bias of estimate [3,4].

Spatial interpolation methods, on the other hand, use information from different sites other than the target station with missing data have also been developed. These methods consider the spatial variability of the measured variable, ignoring the temporal information in long-time series, such methods include the closest station method, the simple arithmetic averaging method; the inverse distance weighted (IDW) method, the single best estimator method and the normal ratio method. These methods generally under and/or overestimate the high and low extremes, respectively [6,7,8].

Comparisons among the different spatial interpolation methods had been carried out in different areas. Most of the work, however, was for areas that were characterized by high rainfall and large number of stations [9,10,11,12]. Similarity index (S index), mean absolute error (MAE), and coefficient of correlation (R) were among the tests that were used for assessment of results of interpolation.

In Jordan, several studies have been conducted on rainfall variability and climate change impacts on water resources [13,14,15]. The country is located in the eastern part of the Mediterranean region (Fig. 1) and experiences high spatiotemporal variability of rainfall withinter seasonal variations in rainfall distribution.The country is suffering from water scarcity which resulted in the over abstraction of groundwater, particularly for irrigation in the highlands areas [16].

Since most of Jordan's land is arid, known as Badia, and receive rainfall that is less than 200 mm, adoption of water harvesting techniques is crucial to develop water and land resources in these zones [17]. This requires good spatial distribution of rainfall gauges which is currently unavailable. Little research has been conducted on rainfall data filling under scarce conditions in Jordan. Such research in these arid areas is important from agricultural, hydrological, and groundwater recharge view [18,19]. The aim of this study is to select the best method to fill rainfall monthly series in an arid area in Jordan (Fig. 1) using monthly rainfall data that extends over 30 years (1980-2010) for three rainfall gauge stations.

II. STUDY AREA

The study area is located within Amman-Zarqa basin north of Jordan and has a total area 1000 km2 (Fig. 1). The study area has cold winters and hot summers. Rainfall season extends from October to May, with most rainfall occurring during December-February. Generally, the area is characterized by low amounts of precipitation and high amounts of evaporation, with frequent droughts and high levels of desertification [20,21].

Competition on groundwater sources among agricultural, domestic and industrial sectors in the study area is high [16]. Therefore, augmentation of water supply in this area shall be prioritized. This requires hydrological studies with credible data for rainfall, which in turn suffers from missing records. A previous study was carried out in the same basin to estimate rainfall from different remote sensing data. Results showed inconsistent results for monthly data, with overestimation for months with high rainfall, and weak correlation for daily data [18].

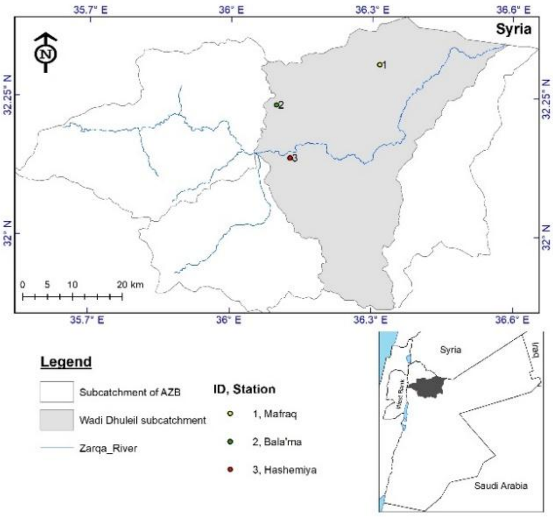

Figure 1: Study area location in north Jordan.

III. MATERIALS AND METHODS

In this study, data for three rainfall gauge stations for the period 1980-2010 was used for filling missing records of rainfall for Wadi Dhuleil subcatchments within Amman-Zarqa Basin (AZB). The stations were located in Mafraq, Hashemiya and Bala'ma. The missing data ranges from few months up to six years (72 months). The rainfall gauge density in this catchment is low (1 gauge per 420 km2) with missing rainfall data ranging between 3% to 12%. There are other stations inside the catchment, but their data suffers from discontinuity and their time series is less than 30 years [18].

3.1 **Estimation methods**

The methods used for estimating the missing rainfall records included the IDW and EM methods.

1- IDW method

In the inverse distance weighted method (IDW), weights for each sample are inversely proportionate to its distance from the point being estimated. The equation of IDW for estimated missing value of rainfall is calculated as follows [22]:

$$Re_i = \frac{\sum_{i=1}^{m} \frac{1}{d_i^n} Ro_i}{\sum_{i=1}^{m} \frac{1}{d_i^n}} \qquad (1)$$

Where,

ReRe_i= estimate of rainfall for the ungauged station,
Ro_i= the rainfall values of rain gauges used for estimation,
d_i= distance from each location to the point being estimated,
m= number of surrounding stations, and
n = user-defined exponent.

The method of IDW can be used for spatial interpolation of any missing parameter [23]. The value of n was set to 2 to give more weight for distance as the high value of the power on the assumption that closer stations are better correlated than those farther away [22].

2- EM method

The expectation maximization method (EM) is an iterative method [24], was proposed based on the reciprocal dependence between the model parameters and the missing values. it can be used to find Maximum Likelihood Estimates for missing data problems.

The EM algorithm consists of two main steps; conditional expectation step and maximization step. The conditional expectations of missing data and estimates of model parameters are calculated by expectation step equation. maximization step finds the estimates of the model parameter to maximize complete data log likelihood function from expectation step. These steps are iterated until the iterations converge [2]. The EM algorithm alternates the expectations and maximization steps for updating the estimate of the unknown parameters at iteration.

3.2**Assessment of results**

There are several performance measures for assessing outputs from estimation or interpolation methods. Most of these are statistical that calculate similarity and errors. The following statistical tests and methods were used for assessing results of estimation.

1- The root means square error (RMSE)

The root mean square error (RMSE) employed in model evaluation studies [25]. RMSE is used to compare the different estimating techniques or methods for identification of the best method. The method with the lowest value of RMSE indicates the best method. The measurement formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Re_i - Roi)^2} \qquad (2)$$

2- Bias or mean of error (ME)

Bias indicates the difference between the estimated value and the real observed value of the parameter. If this result is 0, it indicates unbiased estimation, with the method having the minimum ME being the accurate one [26]. The bias is calculated as follows:

$$\text{Bias} = ME = \frac{1}{n}\sum_{i=1}^{n}(Re_i - \overline{Ro}) \qquad (3)$$

3- The similarity index (S-index)

The similarity index (S-index) is the index of agreement for assessing model performance which implies the percentage of agreement between the observed and estimated values. The values of S-index range from 0.0 for complete disagreement to 1.0 for perfect agreement. S-index formula is as follows:

$$S = 1 - \left[\frac{\sum_{i=1}^{n}(Re - Ro_i)^2}{\sum_{i=1}^{n}(|Re_i - \overline{Ro}| + |Re_i - \overline{Ro}|)^2}\right] \qquad (4)$$

3- Pearson's correlation (R)

Pearson correlation coefficient is a measure of strength and direction of linear relationship between two continous variables *Ro* and *Re*. It has a value between +1 (poisitive linear) and −1 (negative linear), the zero value indicates no correlation. The equation to calculate R is:

$$R = \frac{\sum_{i=1}^{n}(Ro_i - \overline{Ro})(Re_i - \overline{Re})}{\sqrt{\sum_{i=1}^{n}(Ro_i - \overline{Ro})2\ \sum_{i=1}^{n}(Re_i - \overline{Re})^2}} \qquad (5)$$

Where$\overline{Ro}$ is the average of all observed values and $\overline{Re}$ is the average of all estimated values.

IV. RESULTS AND DISCUSSIONS

A summary for the data for the three stations is shown in Table 1. The rainfall datasets showed variations in the missing data percentage for the period October 1980 - May 2010. The correlation test evaluated among stations, based on the original observe data, revealed that the correlation was moderate to good as shown in Table 2.

Table 1: Descriptive statistics for the stations.

Station	Missing %	Mean	Std. Deviation
Bala'ma (A)	5.16	25.71	32.61
Mafraq (B)	2.78	14.72	16.85
Hashemiya (C)	11.95	14.00	18.47

Table 2: Correlation coefficient (R) between the raw data of the stations.

Station	A	B	C
A	1.00		
B	0.77**	1.00	
C	0.80**	0.76**	1.00

** Significant at P <0.01 (2-tailed).

Results for the IDW and EM methods showed that the RMSE did not exceed 15mm, while ME ranged from 0.2 to 4.5 with moderate variations. In general, the RMSE decreased when the missing data percentage decrease. A summary for the statistical tests for the estimation method for the three stations is given in Table 3.

Table 3: Result of evaluation analysis methods.
Station_ Method

Station_ Method	RMSE	S	ME	R
A				
A_EM	8.08	1.00	1.34	1.00**
A_IDW	8.20	1.00	3.97	.982**
B				
B_EM	1.53	1.00	0.25	1.00**
B_IDW	2.58	1.00	1.13	.991**
C				
C_EM	10.65	1.00	3.12	1.00**
C_IDW	13.49	1.00	4.04	.995**

** ** Significant at P < 0.01 (2-tailed).

The resulted RMSE and ME for EM method were lower than those for the IDW for the three stations, while the correlation and the S-index were the higher for EM than for the IDW for the three stations which indicates that EM method is more accurate than IDW for estimating the missing rainfall values. This could be attributed to the nature of EM that estimates the means, the covariance matrix, and the correlation of quantitative variables with missing values using the iterative process and makes inferences based on the likelihood under the specified distribution.

In terms of rainfall stations, variations in accuracy of estimation was observed. Generally, the station with low percent of missing data, Mafraq station in this case, had attained higher accuracy of estimation when compared with the stations with high percent of missing data. Also, IDW method have lower accuracy than EM method as it would follow the trend of correlation (Table 2) and would be more biased to location of the station, showing that it would need higher number of stations than EM method for estimating missing data.

Results of estimation plotted for the time series are shown in Figures 2 to 4. The figures demonstrated the trends of the rainfall time series for both actual and estimated data.

Generally, results from the time series plots of estimated and actual data were compatible with the RMSE and biasness for estimates for the three stations were the value of ME was, for EM method, close to zero showing little biasness for the IDW method.

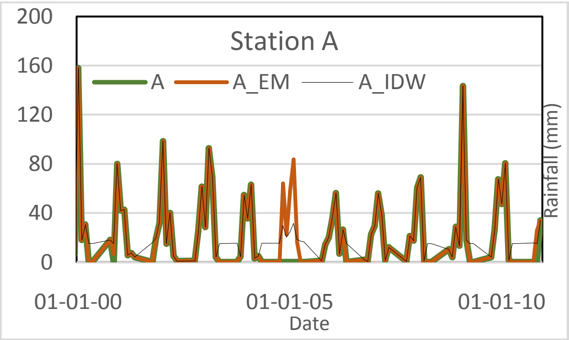

Figure 2: Time series plots of the original and the estimated rainfall series for Bala'ma station.

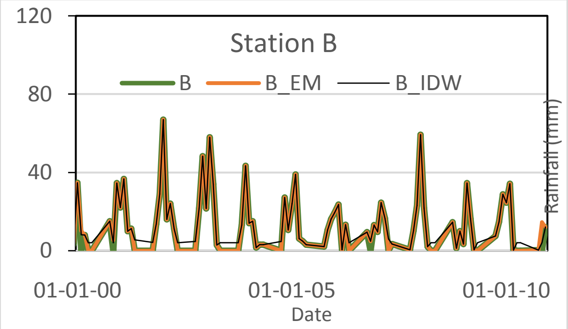

Figure 3: Time series plots of the original and the estimated rainfall series for Mafraq station.

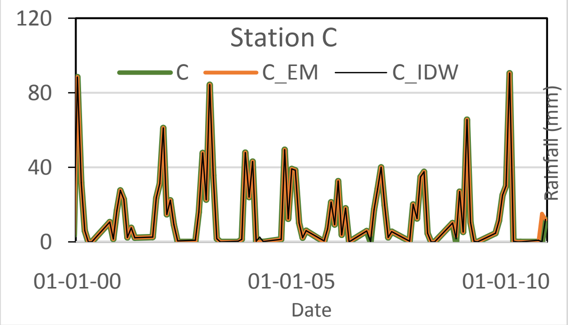

Figure 4: Time series plots of the original and the estimated rainfall series for Hashemiya station.

It is worthy to mention that results from this study showed lower RMSE for estimates and lower R for correlation among station than values reported in literature [9]. This could be attributed to the relatively low levels of rainfall in this arid area when compared with data reported for wet environments. Also, it could be attributed to the low density of stations in the study area compared to other ones from the literature.

Comparing the results obtained from this study with the use of remote sensing data for filling the missing data for the same basin and catchment [18], results obtained from this study showed better estimates particularly for the area closer to Mafraq station. This would suggest the use of EM method for filling the missing monthly rainfall data.

V. CONCLUSIONS

This study focuses on filling the missing monthly rainfall data using EM and IDW methods. The optimal method was the EM as indicated by the different statistical tests. Hence, it can be concluded that in arid areas with different missing values up to 15%, the EM method shows robust results than spatial method, with clear insignificant bias values, however, Adoption of either method might require the inclusion of other methods for estimation and the inclusion of other rainfall stations in the estimation, although the records might be for shorter time series than the one used in this study.

VI. REFERENCES

[1] M Villazón, P. Willems, Filling gaps and Daily Disaccumulation of Precipitation Data for Rainfall-runoff model. Proc.4th International Scientific Conference on Water Observation and Information Systems for Decision Support, Ohrid, Republic of Macedonia, 25–29 May 2010, 1–9.

[2] T Schneider, Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values. Journal of Climate, 14, 2001, 853–871.

[3] M Ben Aissia, F. Chebana, T.B.M.J. Ouarda, Multivariate missing data in hydrology – Review and applications. Advances in Water Resources, 110, 2017, 299–309.

[4] F Jahan, N.C. Sinha, M.M. Rahman, M.S.H. Mondal, M.A. Islam, Comparison of missing value estimation techniques in rainfall data of Bangladesh. Theoretical and Applied Climatology, 2018, 1–17.

[5] H El Sharif, E. Teegavarapu. Evaluation of Spatial Interpolation Methods for Missing Precipitation Data: Preservation of Spatial Statistics; Proceedings; 2018; ISBN 978-0-7844-1231-2.

[6] S K Regonda, D.J. Seo, B. Lawrence, J.D. Brown, J. Demargne, Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts - A Hydrologic Model Output Statistics (HMOS) approach. Journal of Hydrology, 497, 2013, 80–96.

[7] X Jun, Y.D. Chen, Water problems and opportunities in the hydrological sciences in China. Hydrological Sciences Journal, 46, 2001, 907–921.

[8] R S V Teegavarapu, Statistical corrections of spatially interpolated missing precipitation data estimates. Hydrological Processes, 28, 2013, 3789–3808.

[9] R P DeSilva, N.D.K Dayawansa, M.D. Ratnasiri, A comparison of methods used in estimating missing rainfall data. The Journal of Agricultural Science, 3, 2007, 101–108.

[10] J Suhalia, M.D. Sayang, A.A. Jemain, Revised Spatial Weighting Methods for Estimation of Missing Rainfall Data. Asia-Pacific Journal of Atmospheric Sciences, 44, 2008, 93–104.

[11] B Ahrens, Distance in spatial interpolation of daily rain gauge data. Hydrology and Earth System Sciences, 10, 2006, 197–208.

[12] F W Chen, C.W. Liu, Estimation of the spatial rainfall distribution using inverse distance

weighting (IDW) in the middle of Taiwan. Paddy and Water Environment, 10, 2012, 209–222.

[13] J T Al-Bakri, M. Salahat, A. Suleiman, M. Suifan, M.R. Hamdan, S. Khresat, T. Kandakji, Impact of Climate and Land Use Changes on Water and Food Security in Jordan: Implications for Transcending "The Tragedy of the Commons", Sustainability, 5(2), 2013, 724-748.

[14] Q. Tarawneh, M. Kadıoğlu, An analysis of precipitation climatology in Jordan. Theor. Appl. Climatol., 74, 2003, 123–136.

[15] M Freiwan, M. Kadioglu, Spatial and temporal analysis of climatological data in Jordan. Int. J. Climatol., 28, 2007, 521–535.

[16] J T Al-Bakri, S. Shawash, A. Ghanim, R. Abdelkhaleq. Geospatial Techniques for Improved Water Management in Jordan. Water, 8(4), 2016, 132. doi:10.3390/w8040132

[17] F Ziadat, T. Oweis, S. Mazahreh, A. Bruggeman, N. Haddad, E. Karablieh, B. Benli, M. Abu Zanat, J. Al-Bakri, A. Ali, Selection and characterization of Badia watershed research sites (International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria, 2006), vi+111.

[18] E. Abushandi, B. Merkel, Rainfall estimation over the Wadi Dhuliel arid catchment, Jordan from GSMaP_MVK+. Hydrology and Earth System Sciences Discussions, 8, 2011, 1665–1704.

[19] N S Lam, Spatial Interpolation Methods: A Review. Cartography and Geographic Information Science, 10, 1983, 129–150.

[20] J T Al-Bakri, L. Brown, Z. Gedalof, A. Berg, W. Nickling, S. Khresat, M. Salahat H. Saoub. Modelling desertification risk in the north-west of Jordan using geospatial and remote sensing techniques. Geomatics, Natural Hazards and Risk, 7(2), 2016, 531-549.

[21] J T Al-Bakri, A. Al-Khreisat, S. Shawash, E. Qaryouti, M. Saba, Assessment of Remote Sensing Indices for Drought Monitoring in Jordan. Asian Journal of Geoinformatics, 17(3), 2017, 1-13.

[22] P Bolstad, 2016. GIS Fundamentals: A First Text on Geographic Information Systems, 5th Ed. (Eider Press, White Bear Lake, Minnesota).

[23] J T Al-Bakri, D. Al-Eisawi, S. Damhoureyeh, S Oran, GIS-based analysis of spatial distribution of medicinal and herbal plants in the northwest of Jordan. Annals of Arid Zone, 50(2), 2011, 99-115.

[24] A P Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B Methodological, 39, 1977, 1–38.

[25] C J Willmott, On the validation of models, Physical Geography,2, 1981, 184–194.

[26] Shepard, D. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 23rd ACM national conference on - 1968, 517–524.

Zeinab Abu Romman" Estimation of Rainfall missing data in arid area using Spatial and EM methods" International Journal of Engineering Research and Applications (IJERA),Vol. 09, No.03, 2019, pp. 76-80